PCT/AU98/00380

Patent Office
Canberra

I, KIM MARSHALL, MANAGER EXAMINATION SUPPORT AND SALES, hereby certify that the annexed is a true copy of the Provisional specification in connection with Application No. PP 1458 for a patent by THE COUNCIL OF THE QUEENSLAND INSTITUTE OF MEDICAL RESEARCH filed on 22 January 1998.

I further certify that the annexed specification is not, as yet, open to public inspection.

PRIORITY DOCUMENT

WITNESS my hand this First
day of June 1998

KIM MARSHALL
MANAGER EXAMINATION SUPPORT AND SALES

Regulation 3.2

The Council of The Queensland Institute of Medical Research

A U S T R A L I A

Patents Act 1990

PROVISIONAL SPECIFICATION

for the invention entitled:

"A novel gene and uses therefor-IIa"

The invention is described in the following statement:

A NOVEL GENE AND USES THEREFOR-IIa

FIELD OF THE INVENTION

The present invention relates generally to a novel human gene and to derivatives and mammalian, animal, insect, nematodes, avian and microbial homologues thereof. The present invention further provides pharmaceutical compositions and diagnostic agents as well as genetic molecules useful in gene replacement therapy and recombinant molecules useful in protein replacement therapy.

Bibliographic details of the publications referred to by author in this specification are collected at the end of the description. Sequence identity numbers (SEQ ID NOs.) for nucleotide and amino acid sequences referred to in the subject specification are defined after the bibliography.

BACKGROUND OF THE INVENTION

The increasing sophistication of recombinant DNA technology is greatly facilitating research and development in the medical and allied health fields. There is growing need to develop recombinant and genetic molecules for use in diagnosis, conventional pharmaceutical preparations as well as gene and protein replacement therapies.

In work leading up to the present invention, the inventors sought to identify and clone human genes which might be useful as potential diagnostic and/or therapeutic agents. One area of particular interest is in the field of gene regulators.

Gene expression generally requires interaction between a regulatory protein and an appropriate recognition sequence of a target gene. Regulatory proteins comprise in many cases a domain or motif that facilitates binding to DNA. One particular motif comprises small sequence units repeated in tandem with each unit folded about a zinc atom to form separate structural domains. This motif is now referred to as a zinc finger domain. Such a

domain is generally defined by the number of cysteine (C) and histidine (H) residues.

In accordance with the present invention, a gene has been identified from the human genome with an N-terminal region resembling a zinc-finger domain of a novel type.

SUMMARY OF THE INVENTION

Throughout this specification, unless the context requires otherwise, the word "comprise", or variations such as "comprises" or "comprising", will be understood to imply the inclusion of a stated element or integer or group of elements or integers but not the exclusion of any other element or integer or group of elements or integers.

One aspect of the present invention contemplates an isolated nucleic acid molecule comprising a sequence of nucleotides encoding or complementary to a sequence encoding an amino acid sequence having homology to a regulator of gene expression or a derivative of said gene regulator.

Another aspect of the present invention provides an isolated nucleic acid molecule comprising a sequence of nucleotides encoding or complementary to a sequence encoding putative regulator of gene expression wherein said regulator comprises a zinc finger domain of an $(HC_3)_2$ type.

Yet another aspect of the present invention is directed to an isolated nucleic acid molecule comprising a sequence of nucleotides or a complementary form thereof selected from:

(i) a nucleotide sequence set forth in SEQ ID NO:1;

(ii) a nucleotide sequence encoding an amino acid sequence set forth in SEQ ID NO:2;

(iii) a nucleotide sequence having at least about 40% similarity to the nucleotide sequence of (i) or (ii); and

(iv) a nucleotide sequence capable of hybridizing under low stringency conditions to

the nucleotide sequence set forth in (i), (ii) or (iii).

Even yet another aspect of the present invention provides a genetic construct comprising a vector portion and an animal, more particularly a mammalian and even more particularly a human *mcg4* gene portion, which *mcg4* gene portion is capable of encoding an MCG4 polypeptide or a functional or immunologically interactive derivative thereof.

Still yet another aspect of the present invention contemplates a method of detecting a condition caused or facilitated by an aberration in *mcg4,* said method comprising determining the presence of a single or multiple nucleotide substitution, deletion and/or addition or other aberration to one or both alleles of said *mcg4* wherein the presence of such a nucleotide substitution, deletion and/or addition or other aberration may be indicative of said condition or a propensity to develop said condition.

Even still a further aspect of the present invention relates to a method of detecting a condition caused or facilitated by an aberration in *mcg4,* said method comprising screening for a single or multiple amino acid substitution, deletion and/or addition to MCG4 wherein the presence of such a mutation is indicative of or a propensity to develop said condition.

Another aspect of the present invention contemplates a method for detecting MCG4 or a derivative thereof in a biological sample said method comprising contacting said biological sample with an antibody specific for MCG4 or its derivatives or homologues for a time and under conditions sufficient for an antibody-MCG4 complex to form, and then detecting said complex.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is a representation of the nucleotide sequence and corresponding amino acid sequence of *mcg4*.

Figure 2 is a representation of the alignment of the human MCG4 amino acid sequence with a translation of a partial murine expressed sequence tag (EST).

Figure 3 is a representation of the alignment of the human MCG4 amino acid sequence with a translation of a partial nematode EST.

Figure 4 is a diagrammatic representation showing a predicted structure of MCG4 where H and C represent histidine and cysteine residues, respectively and X refers to any amino acid residue. Zn represent zinc atoms.

Figure 5 is a representation of sensitive sequence homology search of related cysteine-containing motifs in another *Caenorhabditis elegans* protein.

Figure 6 is a representation showing that a related cysteine containing motif is present in the GATA-binding transcription factor from *Saccharomyces pombe*.

Figure 7 is a Northern blot showing expression of *mcg4* in various cultured human cancer cell lines. Lanes 1-5, respectively, represent the hybridization signal from 15μg total RNA derived from various human cancer cell lines. Lanes 1-5, respectively, contain RNA from H69 lung carcinoma cells, JAM ovary carcinoma cells, BT20 breast carcinoma cells, HaCat transformed keratinocytes, T24 bladder carcinoma cells.

Figure 8 is a representation of a partial alignment of *mcg4* with human ESTs AA074703 and AA134788.

Figure 9 is a representation of the partial nucleotide sequence alignment between a human (W32939) and mouse (AA242159) *mcg4*-like EST in the putative 5′ UTR of the *mcg4* cDNA. The putative initiation codon is underlined and the region upstream represents 5′ UTR.

Figure 10 is a representation showing MacVector alignment of MCG4 with forward translations of ESTs AA134788 and AA074703. The nucleotide sequences are shown in Figure 8.

Figure 11 is a diagrammatic representation of the domains of MCG4

zinc finger consensus: $CX_2HX_4CX_2CX_4HX_2CX_{17}CX_2CX_{18}HX_2CX_{18}CX_2C$

acidic domain consensus: 9/34 amino acids negatively charged, 0/34 positively charged

basic domain consensus: 13/55 amino acids positively charged, 0/55 negatively charged

leucine zipper domain consensus: $LX_6LX_6RX_6LX_6L$

alternate "novel" leucine zipper-life motif where leucine would not be aligned along the one surface of an alpha helix domain: (aa261) $LX_6LXLX_6LXLX_6L$ (aa 286).

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The present invention provides an isolated nucleic acid molecule comprising a sequence of nucleotides encoding or complementary to a sequence encoding an amino acid sequence having homology to a regulator of gene expression or a derivative of said gene regulator.

More particularly, the present invention is directed to an isolated nucleic acid molecule comprising a sequence of nucleotides encoding or complementary to a sequence encoding putative regulator of gene expression wherein said regulator comprises a zinc finger domain of an $(HC_3)_2$ type.

Still more particularly, the present invention provides an isolated nucleic acid molecule comprising a sequence of nucleotides or a complementary form thereof selected from:

(i) a nucleotide sequence set forth in SEQ ID NO:1;

(ii) a nucleotide sequence encoding an amino acid sequence set forth in SEQ ID NO:2;

(iii) a nucleotide sequence having at least about 40% similarity to the nucleotide sequence of (i) or (ii); and

(iv) a nucleotide sequence capable of hybridizing under low stringency conditions to the nucleotide sequence set forth in (i), (ii) or (iii).

Preferably, the percentage similarity is at least about 50%. More preferably, the percentage similarity is at least about 60%.

Reference herein to a low stringency at 42°C includes and encompasses from at least about 1% v/v to at least about 15% v/v formamide and from at least about 1M to at least about 2M salt for hybridisation, and at least about 1M to at least about 2M salt for washing conditions. Alternative stringency conditions may be applied where necessary, such as medium stringency, which includes and encompasses from at least about 16% v/v to at least about 30% v/v formamide and from at least about 0.5M to at least about 0.9M salt for hybridisation, and at least about 0.5M to at least about 0.9M salt for washing conditions, or high stringency, which includes and encompasses from at least about 31% v/v to at least about 50% v/v formamide and from at least about 0.01M to at least about 0.15M salt for hybridisation, and at least about 0.01M to at least about 0.15M salt for washing conditions.

The term "similarity" as used herein includes exact identity between compared sequences at the nucleotide or amino acid level. Where there is non-identity at the nucleotide level, "similarity" includes differences between sequences which result in different amino acids that are nevertheless related to each other at the structural, functional, biochemical and/or conformational levels. Where there is non-identity at the amino acid level, "similarity" includes amino acids that are nevertheless related to each other at the structural, functional, biochemical and/or conformational levels.

The present invention extends to nucleic acid molecules with percentage similarities of approximately 65%, 70%, 75%, 80%, 85%, 90% or 95% or above or a percentage in between.

The nucleic acid molecule of the present invention is hereinafter referred to as constituting the "*mcg4*" gene. The protein encoded by *mcg4* is referred to herein as "MCG4". The *mcg4*

gene is proposed to encode, in accordance with the present invention, a regulator of gene expression and to comprise the novel zinc finger domain $(HC_3)_2$. A regulator of gene expression includes a transcription factor. Regulation may be at the level of nucleic acid:protein or protein:protein interaction.

The present invention extends to the naturally occurring genomic *mcg4* nucleotide sequence or corresponding cDNA sequence or to derivatives thereof. Derivatives contemplated in the present invention include fragments, parts, portions, mutants, homologues and analogues of MCG4 or the corresponding genetic sequence. Derivatives also include single or multiple amino acid substitutions, deletions and/or additions to MCG4 or single or multiple nucleotide substitutions, deletions and/or additions to *mcg4*. "Additions" to the amino acid or nucleotide sequences include fusions with other peptides, polypeptides or proteins or fusions to nucleotide sequences. Reference herein to "MCG4" or "*mcg4*" includes references to all derivatives thereof including functional derivatives and immunologically interactive derivatives of MCG4.

The *mcg4* of the present invention is particularly exemplified herein from humans and in particular from human chromosome 11q13.

The present invention extends, however, to a range of homologues from, for example, primates, livestock animals (eg. sheep, cows, horses, donkeys, pigs), companion animals (eg. dogs, cats) laboratory test animals (eg. rabbits, mice, rats, guinea pigs), birds (eg. chickens, ducks, geese, parrots), insects, nematodes, eukaryotic microorganisms and captive wild animals (eg. deer, foxes, kangaroos). Reference herein to *mcg4* or MCG4 includes reference to these molecules of human origin as well as novel forms of non-human origin.

The nucleic acid molecules of the present invention may be DNA or RNA. When the nucleic acid molecule is in DNA form, it may be genomic DNA or cDNA. RNA forms of the nucleic acid molecules of the present invention are generally mRNA.

Although the nucleic acid molecules of the present invention are generally in isolated form, they may be integrated into or ligated to or otherwise fused or associated with other genetic molecules such as vector molecules and in particular expression vector molecules. Vectors and expression vectors are generally capable of replication and, if applicable, expression in one or both of a prokaryotic cell or a eukaryotic cell. Preferably, prokaryotic cells include *E. coli, Bacillus sp* and *Pseudomonas sp.* Preferred eukaryotic cells include yeast, fungal, mammalian and insect cells.

Accordingly, another aspect of the present invention contemplates a genetic construct comprising a vector portion and an animal, more particularly a mammalian and even more particularly a human *mcg4* gene portion, which *mcg4* gene portion is capable of encoding an MCG4 polypeptide or a functional or immunologically interactive derivative thereof.

Preferably, the *mcg4* gene portion of the genetic construct is operably linked to a promoter in the vector such that said promoter is capable of directing expression of said *mcg4* gene portion in an appropriate cell.

In addition, the *mcg4* gene portion of the genetic construct may comprise all or part of the gene fused to another genetic sequence such as a nucleotide sequence encoding glutathione-S-transferase or part thereof.

The present invention extends to such genetic constructs and to prokaryotic or eukaryotic cells comprising same.

It is proposed in accordance with the present invention that MCG4 is a transcription factor involved in gene regulation. Mutations in *mcg4* may result in aberrations in gene regulation leading to the development of or a propensity to develop various types of cancer. In this regard, although not wishing to limit the present invention to any one hypothesis or mode of action, it is proposed that *mcg4* or its expression product may be involved in the tissue-specific or temporal regulation of particular genes.

A deletion or aberration in the *mcg4* gene may also be important in the detection of cancer or a propensity to develop cancer. An aberration may be a homozygous mutation or a heterozygous mutation. The detection may occur at the foetal or post-natal level. Detection may also be at the germline or somatic cell level. Furthermore, a risk of developing cancer may be determined by assaying for aberrations in the parents and/or proband of a subject under investigation.

According to this aspect of the present invention, there is contemplated a method of detecting a condition caused or facilitated by an aberration in *mcg4,* said method comprising determining the presence of a single or multiple nucleotide substitution, deletion and/or addition or other aberration to one or both alleles of said *mcg4* wherein the presence of such a nucleotide substitution, deletion and/or addition or other aberration may be indicative of said condition or a propensity to develop said condition.

The nucleotide substitutions, additions or deletions may be detected by any convenient means including nucleotide sequencing, restriction fragment length polymorphism (RFLP), polymerase chain reaction (PCR), oligonucleotide hybridization and single stranded conformation polymorphism analysis (SSCP) amongst many others. An aberration includes modification to existing nucleotides such as to modify glycosylation signal amongst other effects.

In an alternative method, aberrations in the *mcg4* gene are detected by screening for mutations in MCG4.

A mutation in MCG4 may be a single or multiple amino acid substitution, addition and/or deletion. The mutation in *mcg4* may also result in either no translation product being produced or a product in truncated form. A mutant may also be an altered glycosylation pattern or the introduction of side chain modifications to amino acid residues.

According to this aspect of the present invention, there is provided a method of detecting a

condition caused or facilitated by an aberration in *mcg4,* said method comprising screening for a single or multiple amino acid substitution, deletion and/or addition to MCG4 wherein the presence of such a mutation is indicative of or a propensity to develop said condition.

A particularly convenient means of detecting a mutation in MCG4 is by use of antibodies.

Accordingly another aspect of the present invention is directed to antibodies to MCG4 and its derivatives. Such antibodies may be monoclonal or polyclonal and may be selected from naturally occurring antibodies to MCG4 or may be specifically raised to MCG4 or derivatives thereof. In the case of the latter, MCG4 or its derivatives may first need to be associated with a carrier molecule. The antibodies to MCG4 of the present invention are particularly useful as diagnostic agents.

For example, antibodies to MCG4 and its derivatives can be used to screen for wild-type MCG4 or for mutated MCG4 molecules. The latter may occur, for example, during or prior to certain cancer development. A differential binding assay is also particularly useful. Techniques for such assays are well known in the art and include, for example, sandwich assays and ELISA. Knowledge of normal MCG4 levels or the presence of wild-type MCG4 may be important for diagnosis of certain cancers or a predisposition for development of cancers or for monitoring certain therapeutic protocols.

As stated above antibodies to MCG4 of the present invention may be monoclonal or polyclonal or may be fragments of antibodies such as Fab fragments. Furthermore, the present invention extends to recombinant and synthetic antibodies and to antibody hybrids. A "synthetic antibody" is considered herein to include fragments and hybrids of antibodies.

For example, specific antibodies can be used to screen for wild-type MCG4 molecule or specific mutant molecules such as molecules having a certain deletion. This would be important, for example, as a means for screening for levels of MCG4 in a cell extract or other biological fluid or purifying MCG4 made by recombinant means from culture supernatant fluid or purified from

a cell extract. Techniques for the assays contemplated herein are known in the art and include, for example, sandwich assays and ELISA.

It is within the scope of this invention to include any second antibodies (monoclonal, polyclonal or fragments of antibodies or synthetic antibodies) directed to the first mentioned antibodies discussed above. Both the first and second antibodies may be used in detection assays or a first antibody may be used with a commercially available anti-immunoglobulin antibody. An antibody as contemplated herein includes any antibody specific to any region of wild-type MCG4 or to a specific mutant phenotype or to a deleted or otherwise altered region.

Both polyclonal and monoclonal antibodies are obtainable by immunization of a suitable animal or bird with MCG4 or its derivatives and either type is utilizable for immunoassays. The methods of obtaining both types of sera are well known in the art. Polyclonal sera are less preferred but are relatively easily prepared by injection of a suitable laboratory animal or bird with an effective amount of MCG4 or antigenic parts thereof or derivatives thereof, collecting serum from the animal or bird, and isolating specific sera by any of the known immunoadsorbent techniques. Although antibodies produced by this method are utilizable in virtually any type of immunoassay, they are generally less favoured because of the potential heterogeneity of the product.

The use of monoclonal antibodies in an immunoassay is particularly preferred because of the ability to produce them in large quantities and the homogeneity of the product. The preparation of hybridoma cell lines for monoclonal antibody production derived by fusing an immortal cell line and lymphocytes sensitized against the immunogenic preparation can be done by techniques which are well known to those who are skilled in the art.

Another aspect of the present invention contemplates a method for detecting MCG4 or a derivative thereof in a biological sample said method comprising contacting said biological sample with an antibody specific for MCG4 or its derivatives or homologues for a time and under conditions sufficient for an antibody-MCG4 complex to form, and then detecting said

complex.

Preferably, the biological sample is a cell extract from a human or other animal or a bird.

The presence of MCG4 may be accomplished in a number of ways such as by Western blotting and ELISA procedures. A wide range of immunoassay techniques are available as can be seen by reference to US Patent Nos. 4,016,043, 4, 424,279 and 4,018,653. These include both single-site and two-site or "sandwich" assays of the non-competitive types, as well as traditional competitive binding assays. These assays also include direct binding of a labelled antibody to a target.

Sandwich assays are among the most useful and commonly used assays and are favoured for use in the present invention. A number of variations of the sandwich assay technique exist, and all are intended to be encompassed by the present invention. Briefly, in a typical forward assay, an unlabelled antibody is immobilized on a solid substrate and the sample to be tested brought into contact with the bound molecule. After a suitable period of incubation, for a period of time sufficient to allow formation of an antibody-antigen complex, a second antibody specific to the antigen, labelled with a reporter molecule capable of producing a detectable signal is then added and incubated, allowing time sufficient for the formation of another complex of antibody-antigen-labelled antibody. Any unreacted material is washed away, and the presence of the antigen is determined by observation of a signal produced by the reporter molecule. The results may either be qualitative, by simple observation of the visible signal, or may be quantitated by comparing with a control sample containing known amounts of hapten. Variations on the forward assay include a simultaneous assay, in which both sample and labelled antibody are added simultaneously to the bound antibody. These techniques are well known to those skilled in the art, including any minor variations as will be readily apparent. In accordance with the present invention the sample is one which might contain MCG4 including cell extract or, tissue biopsy. The sample is, therefore, generally a biological sample comprising biological fluid but also extends to fermentation fluid and supernatant fluid such as from a cell culture.

In the typical forward sandwich assay, a first antibody having specificity for the MCG4 or an antigenic part thereof or a derivative thereof or antigenic parts thereof, is either covalently or passively bound to a solid surface. The solid surface is typically glass or a polymer, the most commonly used polymers being cellulose, polyacrylamide, nylon, polystyrene, polyvinyl chloride or polypropylene. The solid supports may be in the form of tubes, beads, discs of microplates, or any other surface suitable for conducting an immunoassay. The binding processes are well-known in the art and generally consist of cross-linking covalently binding or physically adsorbing, the polymer-antibody complex is washed in preparation for the test sample. An aliquot of the sample to be tested is then added to the solid phase complex and incubated for a period of time sufficient (e.g. 2-40 minutes) and under suitable conditions (e.g. 25°C) to allow binding of any subunit present in the antibody. Following the incubation period, the antibody subunit solid phase is washed and dried and incubated with a second antibody specific for a portion of the hapten. The second antibody is linked to a reporter molecule which is used to indicate the binding of the second antibody to the hapten.

An alternative method involves immobilizing the target molecules in the biological sample and then exposing the immobilized target to specific antibody which may or may not be labelled with a reporter molecule. Depending on the amount of target and the strength of the reporter molecule signal, a bound target may be detectable by direct labelling with the antibody. Alternatively, a second labelled antibody, specific to the first antibody is exposed to the target-first antibody complex to form a target-first antibody-second antibody tertiary complex. The complex is detected by the signal emitted by the reporter molecule.

By "reporter molecule" as used in the present specification, is meant a molecule which, by its chemical nature, provides an analytically identifiable signal which allows the detection of antigen-bound antibody. Detection may be either qualitative or quantitative. The most commonly used reporter molecules in this type of assay are either enzymes, fluorophores or radionuclide containing molecules (i.e. radioisotopes) and chemiluminescent molecules.

In the case of an enzyme immunoassay, an enzyme is conjugated to the second antibody, generally by means of glutaraldehyde or periodate. As will be readily recognized, however, a

wide variety of different conjugation techniques exist, which are readily available to the skilled artisan. Commonly used enzymes include horseradish peroxidase, glucose oxidase, beta-galactosidase and alkaline phosphatase, amongst others. The substrates to be used with the specific enzymes are generally chosen for the production, upon hydrolysis by the corresponding enzyme, of a detectable colour change. Examples of suitable enzymes include alkaline phosphatase and peroxidase. It is also possible to employ fluorogenic substrates, which yield a fluorescent product rather than the chromogenic substrates noted above. In all cases, the enzyme-labelled antibody is added to the first antibody hapten complex, allowed to bind, and then the excess reagent is washed away. A solution containing the appropriate substrate is then added to the complex of antibody-antigen-antibody. The substrate will react with the enzyme linked to the second antibody, giving a qualitative visual signal, which may be further quantitated, usually spectrophotometrically, to give an indication of the amount of hapten which was present in the sample. "Reporter molecule" also extends to use of cell agglutination or inhibition of agglutination such as red blood cells on latex beads, and the like.

Alternately, fluorescent compounds, such as fluorescein and rhodamine, may be chemically coupled to antibodies without altering their binding capacity. When activated by illumination with light of a particular wavelength, the fluorochrome-labelled antibody adsorbs the light energy, inducing a state to excitability in the molecule, followed by emission of the light at a characteristic colour visually detectable with a light microscope. As in the EIA, the fluorescent labelled antibody is allowed to bind to the first antibody-hapten complex. After washing off the unbound reagent, the remaining tertiary complex is then exposed to the light of the appropriate wavelength the fluorescence observed indicates the presence of the hapten of interest. Immunofluorescence and EIA techniques are both very well established in the art and are particularly preferred for the present method. However, other reporter molecules, such as radioisotope, chemiluminescent or bioluminescent molecules, may also be employed.

As stated above, the present invention extends to genetic constructs capable of encoding MCG4 or functional derivatives thereof. Such genetic constructs are also contemplated to be useful in modulating expression of specific genes in which *mcg4* is involved in tissue-specific

or temporal regulation.

Accordingly, another aspect of the present invention is directed to a genetic construct comprising a nucleotide sequence encoding a peptide, polypeptide or protein and *mcg4* or a functional derivative or homologue thereof capable of modulating the expression of said nucleotide sequence.

The present invention is further described with reference to the following non-limiting Examples.

EXAMPLE 1

A human gene (designated *mcg4*) was identified on chromosome 11q13 that on the basis of sequence homology is predicted to encode a putative transcription factor of 310 amino acids (Fig. 1). *mcg4* is transcribed in several different cell lines (Fig. 7).

EXAMPLE 2

The expressed sequence tag (EST) database contains partial sequence data for the murine (Fig. 2) and nematode (Fig. 3) homologues of *mcg4*.

EXAMPLE 3

MCG4 contains a sequence of cysteine residues within the N-terminal region of the protein that resembles zinc-finger binding domains of a novel type, ie. $(HC_3)_2$ [Fig. 4].

EXAMPLE 4

Sensitive sequence homology searches reveal that related cysteine-containing motifs are present in another *C. elegans* protein (Fig. 5) as well as the GATA-binding transcription factor from *S. pombe* (Fig. 6).

EXAMPLE 5

mcg4 will have commercial value due to its likelihood of encoding a novel transcription factor that is highly conserved amongst organisms, thus suggesting an integral role in gene regulation. *mcg4* may also be involved in some way in tissue-specific or temporal regulation of certain genes, thus making it a potential target for modulating expression of those downstream effectors.

EXAMPLE 6

Nucleotide sequence data generated from cosmid clone cSRL-72c4 with the T7 primer (Promega, and Applied Biosystems Incorporated dye terminator sequencing kit) was aligned to the GenBank Expressed Sequence Tag (EST) database using the program BLASTN (Altschul *et al* 1990) and was found to match numerous human and mouse entries (Table 1 and Figure 2). These matching ESTs were further used to identify overlapping entries in the EST database (Table 1). The nucleotide sequences of these human ESTs were complied using MacVector 4.2.1 software (IBI-Kodak) to produce the cDNA sequence shown in Figure 1. EST entries AA074703 and AA134788 are closely related at the nucleotide level to *mcg4* and it is, therefore, likely that *mcg4* is a member of a newly discovered gene family (Figure 8).

The cDNA sequence of *mcg4* was translated in all possible reading frames and compared to the GenBank non-redundant protein database using the program BLASTX (Altschul *et al* 1990) at the National Center for Biotechnology Information (http//www.ncbi.nih.gov.nlm). As the protein appeared to be novel, a translation of the longest reading frame for the *mcg4* cDNA was aligned to the EST database using the program TBLASTN, which performed a dynamic translation of the EST database in all 6 frames. The search results indicated that the nematode *C. elegans* had an MCG4-like protein (Figure 3), with the matching domains containing a spatial sequence of Cysteine and Histidine residues which resembled a zinc-finger structure (Figure 4). The program BLASTP was used, therefore, to conduct sensitive searches of the protein databases for similar zinc-finger motifs. A weak match to the putative zinc-finger domain was observed for another protein from *C. elegans* (Figure 5) and a poorer match for the GATA-binding transcription factor from *S. pombe* (Figure 6). The putative initiation codon of human *mcg4* is not preceded by an in-frame stop codon and it is therefore possible that the cDNA described in Figure 1 is a truncated form. However, sequence alignment of human and mouse *mcg4* ESTs showed a lower degree of nucleotide conservation prior to the assigned initiation codon, thus supporting the notion that the region represents the 5′ UTR (Figure 9). To determine the expression pattern of *mcg4,* 15μg of the total

cellular RNA (RNeasy Mini Kit, Qiagen) from various human cell lines grown in culture were electrophoresed through 1.2% w/v MOPS/formaldehyde gels and blotted onto nylon membranes (Amersham) by capillary transfer using 20 x SSC (Sambrook *et al* 1989). Filters were subsequently UV-fixed and hybridised overnight at 65°C to a radiolabelled (^{32}P-dCTP) cDNA probe (Church and Gilbert, 1984) for *mcg4*. After washes in 0.1 x SSC/0.1% w/v SDS at 65°C for 1 hour, the filters were air-dried and exposed to X-ray film. This Northern analysis showed that *mcg4* is expressed as a 1.6kb message in numerous tissues including breast, ovary, bladder, lung and keratinocytes (Figure 7).

A MacVector alignment of MCG4 with forward translations of the ESTs AA134788 and AA074703 is shown in Figure 10. The ESTs matching AA074703 are shown in Table 2.

EXAMPLE 7

A diagrammatic representation of the domains of MCG4 is shown in Figure 11.

Those skilled in the art will appreciate that the invention described herein is susceptible to variations and modifications other than those specifically described. It is to be understood that the invention includes all such variations and modifications. The invention also includes all of the steps, features, compositions and compounds referred to or indicated in this specification, individually or collectively, and any and all combinations of any two or more of said steps or features.

TABLE 1

ESTs matching *mcg4*

accession number	seq. run	organism	score	E value	N
gb|AA399110|AA399110	zt89e06.s1	Soares testis NHT Homo sa...	1136	4.0e-168	2
gb|N39612|N39612	yy51g06.s1	Homo sapiens cDNA clone 2...	1521	5.3e-168	4
gb|AA514406|AA514406	nf57d01.s1	NCI_CGAP_Co3 Homo sapiens...	931	5.5e-166	3
gb|AA544946|AA544946	vk38e02.r1	Soares mouse mammary glan...	1207	8.4e-164	2
gb|AA450076|AA450076	zx42a04.s1	Soares total fetus Nb2HF8...	691	2.3e-160	4
gb|AA535731|AA535731	nf88f07.s1	NCI_CGAP_Co3 Homo sapiens...	796	3.5e-158	4
gb|W79710|W79710	zd86f01.r1	Soares fetal heart NbHH19...	1644	1.1e-157	4
gb|AA503531|AA503531	ne47e08.s1	NCI_CGAP_Co3 Homo sapiens...	736	4.0e-156	4
gb|AA450132|AA450132	zx42a04.r1	Soares total fetus Nb2HF8...	1955	3.9e-155	1
gb|AA398068|AA398068	zt89f06.r1	Soares testis NHT Homo sa...	1315	5.4e-148	2
gb|W60405|W60405	zd29h08.r1	Soares fetal heart NbHH19...	1022	1.8e-139	4
gb|W81382|W81382	zd86f01.s1	Soares fetal heart NbHH19...	605	3.5e-125	5
gb|AA047617|AA047617	zf13f07.s1	Soares fetal heart NbHH19...	922	4.6e-125	2
gb|AA282175|AA282175	zt02d03.s1	NCI_CGAP_GCB1 Homo sapien...	1577	2.0e-123	1
gb|AA242159|AA242159	my30d04.r1	Barstead mouse pooled org...	866	7.7e-117	2
gb|AA068680|AA068680	mm61a05.r1	Stratagene mouse embryoni...	1280	1.6e-98	1
gb|W46766|W46766	zc36b07.s1	Soares senescent fibrobla...	506	9.6e-92	3
gb|N93704|N93704	zb51c04.s1	Soares fetal lung NbHL19W...	584	9.0e-91	4
gb|AA155210|AA155210	mr98e01.r1	Stratagene mouse embryoni...	840	7.6e-87	2
gb|AA366022|AA366022	EST76915	Pineal gland II Homo sapien...	1077	2.4e-81	1
gb|AA037691|AA037691	zk34h12.s1	Soares pregnant uterus Nb...	949	2.1e-80	2
gb|W35374|W35374	zc07h03.s1	Soares parathyroid tumor ...	1016	3.1e-76	1
dbj|C00696|C00696	HUMGS000825l,	Human Gene Signature, ...	1009	1.2e-75	1
gb|T98249|T98249	ye59a07.s1	Homo sapiens cDNA clone 1...	998	6.7e-75	1
gb|W21588|W21588	zb51c04.r1	Soares fetal lung NbHL19W...	484	1.1e-69	4
gb|H32171|H32171	EST107015	Rattus sp. cDNA 5' end.	828	1.1e-60	1
gb|AA108092|AA108092	mm89e06.r1	Stratagene mouse embryoni...	782	1.3e-60	2
gb|AA017857|AA017857	mh44d10.r1	Soares mouse placenta 4Nb...	665	2.5e-60	2
gb|AA037690|AA037690	zk34h12.r1	Soares pregnant uterus Nb...	540	9.4e-53	2
gb|AA531006|AA531006	nj07b11.s1	NCI_CGAP_Pr22 Homo sapien...	535	5.4e-48	2
gb|N46760|N46760	yy51g06.r1	Homo sapiens cDNA clone 2...	665	9.5e-47	1
gb|W23584|W23584	zc71d03.s1	Soares fetal heart NbHH19...	457	1.8e-44	2
gb|W42214|W42214	mc69h09.r1	Soares mouse embryo NbME1...	460	1.3e-38	3
gb|AA244877|AA244877	mx25a04.r1	Soares mouse NML Mus musc...	429	2.9e-25	1
gb|W32939|W32939	zc07h03.r1	Soares parathyroid tumor ...	320	4.8e-18	1

Table 2

ESTs matching AA074703 (*mcg*4-related cDNA)

Database: Non-redundant Database of GenBank EST Division
1,222,625 sequences; 449,352,662 total letters.

Sequences producing High-scoring Segment Pairs: accession number	seq. run	organism	High Score score	Smallest Sum Probability P(N) E value	N
gb|AA074703|AA074703	zm76g07.r1	Stratagene neuroepitheli...	2071	4.0e-167	1
gb|AA068680|AA068680	mm61a05.r1	Stratagene mouse embryon...	1270	4.4e-145	4
gb|AA134788|AA134788	zm81g02.r1	Stratagene neuroepitheli...	946	1.3e-144	5
gb|AA399110|AA399110	zt89e06.s1	Soares testis NHT Homo s...	520	8.7e-119	6
gb|N39612|N39612	yy51g06.s1	Homo sapiens cDNA clone ...	582	9.6e-110	7
gb|AA282175|AA282175	zt02d03.s1	NCI_CGAP_GCB1 Homo sapie...	771	9.4e-80	3
gb|W81382|W81382	zd86f01.s1	Soares fetal heart NbHH1...	329	1.6e-75	6
gb|AA544946|AA544946	vk38e02.r1	Soares mouse mammary gla...	644	9.6e-63	2
gb|W35374|W35374	zc07h03.s1	Soares parathyroid tumor...	294	4.5e-42	4
gb|W57106|W57106	md57c12.r1	Soares mouse embryo NbME...	394	1.9e-30	2
gb|AA244877|AA244877	mx25a04.r1	Soares mouse NML Mus mus...	162	2.1e-27	4
gb|AA017857|AA017857	mh44d10.r1	Soares mouse placenta 4N...	230	3.7e-23	3
gb|AA531006|AA531006	nj07b11.s1	NCI_CGAP_Pr22 Homo sapie...	139	2.3e-19	3
gb|H32171|H32171	EST107015	Rattus sp. cDNA 5' end.	207	2.6e-10	2
gb|W79710|W79710	zd86f01.r1	Soares fetal heart NbHH1...	157	0.0073	1

BIBLIOGRAPHY

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol. 215:* 403-410.

2. Church, G., and Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA 18:* 1991-1995.

3. Sambrook, J., Frtisch, E.F., and Maniatis, T. (1989) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbour Laboratory, Cold Spring Harbour, NY, USA.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: The Council of The Queensland Institute of Medical Research

(ii) TITLE OF INVENTION:A NOVEL GENE AND USES THEREFOR

(iii) NUMBER OF SEQUENCES: 2

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: DAVIES COLLISON CAVE
(B) STREET: 1 LITTLE COLLINS STREET
(C) CITY: MELBOURNE
(D) STATE: VICTORIA
(E) COUNTRY: AUSTRALIA
(F) ZIP: 3000

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.25

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: AUSTRALIAN PROVISIONAL
(B) FILING DATE:
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: HUGHES, DR E JOHN L
(C) REFERENCE/DOCKET NUMBER: EJH/AF

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: +61 3 9254 2777
(B) TELEFAX: +61 3 9254 2770
(C) TELEX: AA 31787

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1242 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 30..959

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
TCAGTAAACA CAGAGACTGG GGATCGATC ATG GGG CTT TGT AAG TGC CCC AAG        53
                                 Met Gly Leu Cys Lys Cys Pro Lys
                                   1               5

AGA AAG GTG ACC AAC CTG TTC TGC TTC GAA CAT CGG GTC AAC GTC TGC       101
Arg Lys Val Thr Asn Leu Phe Cys Phe Glu His Arg Val Asn Val Cys
     10                  15                  20

GAG CAC TGC CTG GTA GCC AAT CAC GCC AAG TGC ATC GTC CAG TCC TAC       149
Glu His Cys Leu Val Ala Asn His Ala Lys Cys Ile Val Gln Ser Tyr
 25                  30                  35                  40

CTG CAA TGG CTC CAA GAT AGC GAC TAC AAC CCC AAT TGC CGC CTG TGC       197
Leu Gln Trp Leu Gln Asp Ser Asp Tyr Asn Pro Asn Cys Arg Leu Cys
                 45                  50                  55

AAC ATA CCC CTG GCC AGC CGA GAG ACG ACC CGC CTT GTC TGC TAT GAT       245
Asn Ile Pro Leu Ala Ser Arg Glu Thr Thr Arg Leu Val Cys Tyr Asp
             60                  65                  70

CTC TTT CAC TGG GCC TGC CTC AAT GAA CGT GCT GCC CAG CTA CCC CGA       293
Leu Phe His Trp Ala Cys Leu Asn Glu Arg Ala Ala Gln Leu Pro Arg
         75                  80                  85

AAC ACG GCA CCT GCC GGC TAT CAG TGC CCC AGC TGC AAT GGC CCC ATC       341
Asn Thr Ala Pro Ala Gly Tyr Gln Cys Pro Ser Cys Asn Gly Pro Ile
     90                  95                  100

TTC CCC CCA ACC AAC CTG GCT GGC CCC GTG GCC TCC GCA CTG AGA GAG       389
Phe Pro Pro Thr Asn Leu Ala Gly Pro Val Ala Ser Ala Leu Arg Glu
105                 110                 115                 120

AAG CTG GCC ACA GTC AAC TGG GCC CGG GCA GGA CTG GGC CTC CCT CTG       437
Lys Leu Ala Thr Val Asn Trp Ala Arg Ala Gly Leu Gly Leu Pro Leu
                125                 130                 135
```

```
ATC GAT GAG GTG GTG AGC CCA GAG CCC GAG CCC CTC AAC ACG TCT GAC        485
Ile Asp Glu Val Val Ser Pro Glu Pro Glu Pro Leu Asn Thr Ser Asp
            140                 145                 150

TTC TCT GAC TGG TCT AGT TTT AAT GCC AGC AGT ACC CCT GGA CCA GAG        533
Phe Ser Asp Trp Ser Ser Phe Asn Ala Ser Ser Thr Pro Gly Pro Glu
        155                 160                 165

GAG GTA GAC AGC GCC TCT GCT GCC CCA GCC TTC TAC AGC CGA GCC CCC        581
Glu Val Asp Ser Ala Ser Ala Ala Pro Ala Phe Tyr Ser Arg Ala Pro
    170                 175                 180

CGG CCC CCA GCT TCC CCA GGC CGG CCC GAG CAG CAC ACA GTG ATC CAC        629
Arg Pro Pro Ala Ser Pro Gly Arg Pro Glu Gln His Thr Val Ile His
185                 190                 195                 200

ATG GGC AAT CCT GAG CCC TTG ACT CAC GCC CCT AGG AAG GTG TAT GAT        677
Met Gly Asn Pro Glu Pro Leu Thr His Ala Pro Arg Lys Val Tyr Asp
                205                 210                 215

ACG CGG GAT GAT GAC CGG ACA CCA GGC CTC CAT GGA GAC TGT GAC GAT        725
Thr Arg Asp Asp Asp Arg Thr Pro Gly Leu His Gly Asp Cys Asp Asp
            220                 225                 230

GAC AAG TAC CGA CGT CGG CCG GCC TTG GGT TGG CTG GCC CGG CTG CTA        773
Asp Lys Tyr Arg Arg Arg Pro Ala Leu Gly Trp Leu Ala Arg Leu Leu
        235                 240                 245

AGG AGC CGG GCT GGG TCT CGG AAG CGG CCG CTG ACC CTG CTC CAG CGG        821
Arg Ser Arg Ala Gly Ser Arg Lys Arg Pro Leu Thr Leu Leu Gln Arg
    250                 255                 260

GCG GGG CTG CTG CTA CTC TTG GGA CTG CTG GGC TTC CTG GCC CTC CTT        869
Ala Gly Leu Leu Leu Leu Leu Gly Leu Leu Gly Phe Leu Ala Leu Leu
265                 270                 275                 280

GCC CTC ATG TCT CGC CTA GGC CGG GCC GCA GCT GAC AGC GAT CCC AAC        917
Ala Leu Met Ser Arg Leu Gly Arg Ala Ala Ala Asp Ser Asp Pro Asn
                285                 290                 295

CTG GAC CCA CTC ATG AAC CCT CAC ATC CGC GTG GGC CCC TCC TGA            962
Leu Asp Pro Leu Met Asn Pro His Ile Arg Val Gly Pro Ser   *
            300                 305                 310

GCCCCCTTGC TTGTGGCTAG GCCAGCCTAG GATGTGGGTT CTGTGGAGGA GAGGCGGGGT     1022

AATGGGGAGG CTGAGGGCAC CTCTTCACTG CCCCTCTCCC TCAAGCCTAA GACACTAAGA     1082

CCCCAGACCC AAAGCCAAGT CCACCAGAGT GGCTCGCAGG CCAGGCCTGG AGTCCCCGTG     1142

GGTCAAGCAT TTGTCTTGAC TTGCTTTCTC CCGGGTCTCC AGCCTCCGAC CCCTCGCCCC     1202

ATGAAGGAGC TGGCAGGTGG AAATAAACAA CAACTTTATT                           1242
```

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 310 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Met Gly Leu Cys Lys Cys Pro Lys Arg Lys Val Thr Asn Leu Phe Cys
 1               5                  10                  15

Phe Glu His Arg Val Asn Val Cys Glu His Cys Leu Val Ala Asn His
            20                  25                  30

Ala Lys Cys Ile Val Gln Ser Tyr Leu Gln Trp Leu Gln Asp Ser Asp
        35                  40                  45

Tyr Asn Pro Asn Cys Arg Leu Cys Asn Ile Pro Leu Ala Ser Arg Glu
    50                  55                  60

Thr Thr Arg Leu Val Cys Tyr Asp Leu Phe His Trp Ala Cys Leu Asn
65                  70                  75                  80

Glu Arg Ala Ala Gln Leu Pro Arg Asn Thr Ala Pro Ala Gly Tyr Gln
                85                  90                  95

Cys Pro Ser Cys Asn Gly Pro Ile Phe Pro Pro Thr Asn Leu Ala Gly
            100                 105                 110

Pro Val Ala Ser Ala Leu Arg Glu Lys Leu Ala Thr Val Asn Trp Ala
        115                 120                 125

Arg Ala Gly Leu Gly Leu Pro Leu Ile Asp Glu Val Val Ser Pro Glu
    130                 135                 140

Pro Glu Pro Leu Asn Thr Ser Asp Phe Ser Asp Trp Ser Ser Phe Asn
145                 150                 155                 160

Ala Ser Ser Thr Pro Gly Pro Glu Glu Val Asp Ser Ala Ser Ala Ala
                165                 170                 175

Pro Ala Phe Tyr Ser Gln Ala Pro Arg Pro Pro Ala Ser Pro Gly Arg
            180                 185                 190

Pro Glu Gln His Thr Val Ile His Met Gly Asn Pro Glu Pro Leu Thr
        195                 200                 205

His Ala Pro Arg Lys Val Tyr Asp Thr Arg Asp Asp Asp Arg Thr Pro
    210                 215                 220

Gly Leu His Gly Asp Cys Asp Asp Asp Lys Tyr Arg Arg Arg Pro Ala
```

```
225                 230                 235                 240

Leu Gly Trp Leu Ala Arg Leu Leu Arg Ser Arg Ala Gly Ser Arg Lys
                245                 250                 255

Arg Pro Leu Thr Leu Leu Gln Arg Ala Gly Leu Leu Leu Leu Leu Gly
            260                 265                 270

Leu Leu Gly Phe Leu Ala Leu Leu Ala Leu Met Ser Arg Leu Gly Arg
        275                 280                 285

Ala Ala Ala Asp Ser Asp Pro Asn Leu Asp Pro Leu Met Asn Pro His
    290                 295                 300

Ile Arg Val Gly Pro Ser
305                 310
```

DATED this 22nd day of January, 1998

The Council of The Queensland Institute of Medical Research

By DAVIES COLLISON CAVE

Patent Attorneys for the Applicants

FIGURE 1

```
TCAGTAAACA CAGAGACTGG GGATCGATC ATG GGG CTT TGT AAG TGC CCC AAG          53
                                Met Gly Leu Cys Lys Cys Pro Lys
                                  1               5

AGA AAG GTG ACC AAC CTG TTC TGC TTC GAA CAT CGG GTC AAC GTC TGC         101
Arg Lys Val Thr Asn Leu Phe Cys Phe Glu His Arg Val Asn Val Cys
     10                  15                  20

GAG CAC TGC CTG GTA GCC AAT CAC GCC AAG TGC ATC GTC CAG TCC TAC         149
Glu His Cys Leu Val Ala Asn His Ala Lys Cys Ile Val Gln Ser Tyr
 25                  30                  35                  40

CTG CAA TGG CTC CAA GAT AGC GAC TAC AAC CCC AAT TGC CGC CTG TGC         197
Leu Gln Trp Leu Gln Asp Ser Asp Tyr Asn Pro Asn Cys Arg Leu Cys
                 45                  50                  55

AAC ATA CCC CTG GCC AGC CGA GAG ACG ACC CGC CTT GTC TGC TAT GAT         245
Asn Ile Pro Leu Ala Ser Arg Glu Thr Thr Arg Leu Val Cys Tyr Asp
             60                  65                  70

CTC TTT CAC TGG GCC TGC CTC AAT GAA CGT GCT GCC CAG CTA CCC CGA         293
Leu Phe His Trp Ala Cys Leu Asn Glu Arg Ala Ala Gln Leu Pro Arg
         75                  80                  85

AAC ACG GCA CCT GCC GGC TAT CAG TGC CCC AGC TGC AAT GGC CCC ATC         341
Asn Thr Ala Pro Ala Gly Tyr Gln Cys Pro Ser Cys Asn Gly Pro Ile
     90                  95                 100

TTC CCC CCA ACC AAC CTG GCT GGC CCC GTG GCC TCC GCA CTG AGA GAG         389
Phe Pro Pro Thr Asn Leu Ala Gly Pro Val Ala Ser Ala Leu Arg Glu
105                 110                 115                 120

AAG CTG GCC ACA GTC AAC TGG GCC CGG GCA GGA CTG GGC CTC CCT CTG         437
Lys Leu Ala Thr Val Asn Trp Ala Arg Ala Gly Leu Gly Leu Pro Leu
                125                 130                 135

ATC GAT GAG GTG GTG AGC CCA GAG CCC GAG CCC CTC AAC ACG TCT GAC         485
Ile Asp Glu Val Val Ser Pro Glu Pro Glu Pro Leu Asn Thr Ser Asp
            140                 145                 150

TTC TCT GAC TGG TCT AGT TTT AAT GCC AGC AGT ACC CCT GGA CCA GAG         533
Phe Ser Asp Trp Ser Ser Phe Asn Ala Ser Ser Thr Pro Gly Pro Glu
        155                 160                 165

GAG GTA GAC AGC GCC TCT GCT GCC CCA GCC TTC TAC AGC CGA GCC CCC         581
Glu Val Asp Ser Ala Ser Ala Ala Pro Ala Phe Tyr Ser Arg Ala Pro
    170                 175                 180

CGG CCC CCA GCT TCC CCA GGC CGG CCC GAG CAG CAC ACA GTG ATC CAC         629
Arg Pro Pro Ala Ser Pro Gly Arg Pro Glu Gln His Thr Val Ile His
185                 190                 195                 200

ATG GGC AAT CCT GAG CCC TTG ACT CAC GCC CCT AGG AAG GTG TAT GAT         677
Met Gly Asn Pro Glu Pro Leu Thr His Ala Pro Arg Lys Val Tyr Asp
                205                 210                 215
```

```
ACG CGG GAT GAT GAC CGG ACA CCA GGC CTC CAT GGA GAC TGT GAC GAT        725
Thr Arg Asp Asp Asp Arg Thr Pro Gly Leu His Gly Asp Cys Asp Asp
            220                 225                 230

GAC AAG TAC CGA CGT CGG CCG GCC TTG GGT TGG CTG GCC CGG CTG CTA        773
Asp Lys Tyr Arg Arg Arg Pro Ala Leu Gly Trp Leu Ala Arg Leu Leu
        235                 240                 245

AGG AGC CGG GCT GGG TCT CGG AAG CGG CCG CTG ACC CTG CTC CAG CGG        821
Arg Ser Arg Ala Gly Ser Arg Lys Arg Pro Leu Thr Leu Leu Gln Arg
    250                 255                 260

GCG GGG CTG CTG CTA CTC TTG GGA CTG CTG GGC TTC CTG GCC CTC CTT        869
Ala Gly Leu Leu Leu Leu Leu Gly Leu Leu Gly Phe Leu Ala Leu Leu
265                 270                 275                 280

GCC CTC ATG TCT CGC CTA GGC CGG GCC GCA GCT GAC AGC GAT CCC AAC        917
Ala Leu Met Ser Arg Leu Gly Arg Ala Ala Ala Asp Ser Asp Pro Asn
                285                 290                 295

CTG GAC CCA CTC ATG AAC CCT CAC ATC CGC GTG GGC CCC TCC TGA            962
Leu Asp Pro Leu Met Asn Pro His Ile Arg Val Gly Pro Ser   *
            300                 305                 310

GCCCCCTTGC TTGTGGCTAG GCCAGCCTAG GATGTGGGTT CTGTGGAGGA GAGGCGGGGT    1022

AATGGGGAGG CTGAGGGCAC CTCTTCACTG CCCCTCTCCC TCAAGCCTAA GACACTAAGA    1082

CCCCAGACCC AAAGCCAAGT CCACCAGAGT GGCTCGCAGG CCAGGCCTGG AGTCCCCGTG    1142

GGTCAAGCAT TTGTCTTGAC TTGCTTTCTC CCGGGTCTCC AGCCTCCGAC CCCTCGCCCC    1202

ATGAAGGAGC TGGCAGGTGG AAATAAACAA CAACTTTATT                          1242
```

Figure 2

```
gb|AA155210|AA155210 mr98e01.r1 Stratagene mouse embryonic carcinoma
           (#937317) Mus musculus cDNA clone 605496 5'

Query:      1 MGLCKCPKRKVTNLFCFEHRVNVCEHCLVANHAKCIVQSYLQWLQDSDYNPNCRLCNIPL 60
              MGLCKCPKRKVTNLFCFEHRVNVCEHCLVANHAKCIVQSYLQWLQDSDYNPNCRLCN PL
Sbjct:     98 MGLCKCPKRKVTNLFCFEHRVNVCEHCLVANHAKCIVQSYLQWLQDSDYNPNCRLCNTPL 277
```

Figure 3

```
dbj|D75913|CELK111G3F C.elegans cDNA clone yk111g3 : 5' end, single read.

Query:      7 PKRKVTNLFCFEHRVNVCEHCLVANHAKCIVQSYLQWLQDSDYNPNCRLCNIPLASRETT 66
              PKRKVTNLF +EHRVNVCE  LV NH  C+VQSYL WL D DY+PNC LC    L   +T
Sbjct:      1 PKRKVTNLFXYEHRVNVCELXLVDNHPNCVVQSYLTWLTDQDYDPNCSLCKTTLXEGDTI 180

Query:     67 RLVCYDLFHWACLNERAAQLPRNTAPAGYQCP 98    98 PSCNGPIFPPNQ 109
              RL C  L HW C +E     P  TAP GY+CP         P C+  +FPP+Q
Sbjct:    181 RLNCLHLLHWKCFDEWXGNFPDTTAPXGYRCP 276  275 PCCSQEVFPPDQ 310
```

Figure 4

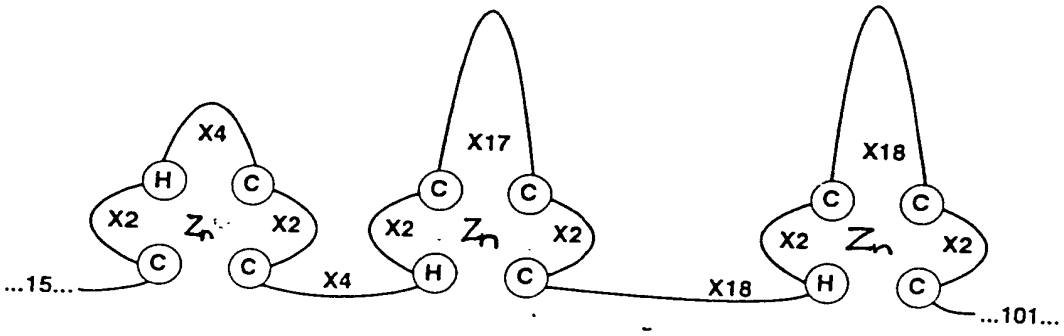


Figure 5

```
sp|P46580|YLB5_CAEEL HYPOTHETICAL 146.8 KD PROTEIN C34E10.5 IN
           CHROMOSOME III gi|500728 (U10402) C34E10.5 gene product
           [Caenorhabditis elegans]

Query:    56 CNIPLASRETTRLVCYDLFHWACLNERAAQLPRNTAPAGYQCPSC 100
             C+I L ++  + L C  LF W C+ E A     + + +  +CP C
Sbjct:  1222 CSICLENKNPSALFCGHLFCWTCIQEHAVAATSSASTSSARCPQC 1266
```

Figure 6

```
gi|703468 (L29051) homologous to GATA-binding transcription factor
           [Schizosaccharomyces pombe]

Query:    35 CIVQSYLQWLQDSDYNPNCRLCNI 58
             C   +  +W +D   NP C  C +
Sbjct:   175 CATTNTPKWRRDESGNPICNACGL 198

Query:   162 SSTPGPEEVDSASAAPAFYSQAPRPPASPGRPEQHTVIHMGNPEPLTHAPRKVYDTRDDD 221
             +S   PEE  S S       S  P+   SP + +Q  +I              P +V  +  D
Sbjct:   441 ASLLNPEEPPSNSDKQPSMSNGPKSEVSPSQSQQAPLIQSSTSPVSLQFPPEVQGSNVDK 500

Query:   222 RTPGLH 227
             R   L+
Sbjct:   501 RNYALN 506
```

Figure 7

Figure 8

```
gb|AA074703|AA074703 zm76g07.r1 Stratagene neuroepithelium (#937231)
           Homo sapiens cDNA clone 531612 5'
           Length = 417

  Plus Strand HSPs:

 Score = 818 (226.0 bits), Expect = 6.1e-103, Sum P(5) = 6.1e-103
 Identities = 206/259 (79%), Positives = 206/259 (79%), Strand = Plus / Plus

Query:   446 GGCCTCCCTCTGATCGATGAGGTGGTGAGCCCAGAGCCCGAGCCCCTCAACACGTCTGAC 505
            || |||||||||||||||||||||| | |||||||||||||||||||||||  | || |||
Sbjct:    49 GGGCTCCCTCTGATCGATGAGGTGATAAGCCCAGAGCCCGAGCCCCTCAATTCCTCAGAC 108

Query:   506 TTCTCTGACTGGTCTAGTTTTAATGCCAGCAGTACCCCTGGACCAGAGGAGGTAGACAGC 565
            |||||||| ||||| || |||||||||| ||  ||| |||  | |||||||  || ||||
Sbjct:   109 TTCTCTGATTGGTCCAGCTTTAATGCCACCACCACCTCTGTGCAAGAGGAGAGAGCCAGC 168

Query:   566 GCCTCTGCTGCCCCAGCCTTCTACAGCCAGGCCCCCCGGCCCCCAGCTTCCCCAGGCCGG 625
              |  |  |||| || || ||||| |||||||| ||||| || ||  | |||||| ||||
Sbjct:   169 ACTCCATCTGCACCTGCTTTCTATAGCCAGGCTCCCCGCCCTCCTCCCTCCCCAAGCCGT 228

Query:   626 CCCGAGCAGCACACAGTGATCCACATGGGCAATCCTGAGCCCTTGACTCACGCCCCTAGG 685
            |||||||||||||||| || |||||||| | | ||||  || || | |||||||| |||
Sbjct:   229 CCCGAGCAGCACACAGTCATACACATGGGGAGTACTGAAGCCCTGGCACACGCCCCAAGG 288

Query:   686 AAGGTGTATGATACGCGGG 704
            || || ||||| || | ||
Sbjct:   289 AAAGTATATGACACACCGG 307

 Score = 230 (63.6 bits), Expect = 6.1e-103, Sum P(5) = 6.1e-103
 Identities = 50/55 (90%), Positives = 50/55 (90%), Strand = Plus / Plus

Query:   398 GCACTGAGAGAGAAGCTGGCCACAGTCAACTGGGCCCGGGCAGGACTGGGCCTCC 452
            ||||||||||| ||||| |||||||||||||| ||||||||||||||||||  ||
Sbjct:     2 GCACTGAGAGAAAAGCTAGCCACAGTCAACTTGGCCCGGGCAGGACTGGGCTCCC 56

 Score = 175 (48.4 bits), Expect = 6.1e-103, Sum P(5) = 6.1e-103
 Identities = 39/44 (88%), Positives = 39/44 (88%), Strand = Plus / Plus

Query:   767 GCCTTGGGTTGGCTGGCCCGGCTGCTAAGGAGCCGGGCTGGGTC 810
            ||  |||| |||||||||| |||||| ||||||||||||||||
Sbjct:   373 GCTCTGGGCTGGCTGGCCCAGCTGCTCAGGAGCCGGGCTGGGTC 416

 Score = 139 (38.4 bits), Expect = 6.1e-103, Sum P(5) = 6.1e-103
 Identities = 31/35 (88%), Positives = 31/35 (88%), Strand = Plus / Plus

Query:   731 GGAGACTGTGACGATGACAAGTACCGACGTCGGCC 765
            |||||||||| |||||||| ||||| || |||||
Sbjct:   336 GGAGACTGTGATGATGACAAATACCGCCGCCGGCC 370

 Score = 133 (36.8 bits), Expect = 6.1e-103, Sum P(5) = 6.1e-103
 Identities = 29/32 (90%), Positives = 29/32 (90%), Strand = Plus / Plus

Query:   701 CGGGATGATGACCGGACACCAGGCCTCCATGG 732
            |||||||||||||||||| ||||| | |||||
Sbjct:   305 CGGGATGATGACCGGACAGCAGGCATTCATGG 336
```

Figure 8 continued

```
gb|AA134788|AA134788 zm81g02.r1 Stratagene neuroepithelium (#937231)
           Homo sapiens cDNA clone 532082 5'
           Length = 368

  Plus Strand HSPs:

 Score = 563 (155.6 bits), Expect = 3.8e-87, Sum P(3) = 3.8e-87
 Identities = 147/190 (77%), Positives = 147/190 (77%), Strand = Plus / Plus

Query:   498 CGTCTGACTTCTCTGACTGGTCTAGTTTTAATGCCAGCAGTACCCCTGGACCAGAGGAGG 557
             | || |||||||||||| ||||| || |||||||||| ||  ||| |||  | |||||||
Sbjct:   103 CCTCAGACTTCTCTGATTGGTCCAGCTTTAATGCCACCACCACCTCTGTGCAAGAGGAGA 162

Query:   558 TAGACAGCGCCTCTGCTGCCCCAGCCTTCTACAGCCAGGCCCCCCGGCCCCCAGCTTCCC 617
              || |||| |  |  |||| || || ||||| |||||||| ||||| || ||  | ||||
Sbjct:   163 GAGCCAGCACTCCATCTGCGCCTGCTTTCTATAGCCAGGCTCCCCGCCCTCCTCCCTCCC 222

Query:   618 CAGGCCGGCCCGAGCAGCACACAGTGATCCACATGGGCAATCCTGAGCCCTTGACTCACG 677
             || |||| |||||||||||||||| || |||||||| | | ||||  || || | ||||
Sbjct:   223 CAAGCCGTCCCGAGCAGCACACAGTCATACACATGGGGAGTACTGAAGCCCTGGCACACG 282

Query:   678 CCCCTAGGAA 687
             |||| |||||
Sbjct:   283 CCCCAAGGAA 292

 Score = 454 (125.4 bits), Expect = 3.8e-87, Sum P(3) = 3.8e-87
 Identities = 94/98 (95%), Positives = 94/98 (95%), Strand = Plus / Plus

Query:   398 GCACTGAGAGAGAAGCTGGCCACAGTCAACTGGGCCCGGGCAGGACTGGGCCTCCCTCTG 457
             |||||||||||| ||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct:     2 GCACTGAGAGACAAGCTAGCCACAGTCAACTGGGCCCGGGCAGGACTGGGCCTCCCTCTG 61

Query:   458 ATCGATGAGGTGGTGAGCCCAGAGCCCGAGCCCCTCAA 495
             |||||||||||| | ||||||||||||||||||||||
Sbjct:    62 ATCGATGAGGTGATAAGCCCAGAGCCCGAGCCCCTCAA 99

 Score = 219 (60.5 bits), Expect = 3.8e-87, Sum P(3) = 3.8e-87
 Identities = 51/60 (85%), Positives = 51/60 (85%), Strand = Plus / Plus

Query:   702 GGGATGATGACCGGACACCAGGCCTCCATGGAGACTGTGACGATGACAAGTACCGACGTC 761
             ||  |||||||||||||| ||||| | |||||||||||||| |||||||| ||||| || |
Sbjct:   309 GGATTGATGACCGGACAGCAGGCATTCATGGAGACTGTGATGATGACAAATACCGCCGCC 368
```

Figure 9

```
W32939    human TACCGCCCTTCGGAACCAGTGCAGCGGCCGATCAGTAAACACAGAGACTGGGGATCGATCATGGGGCTTTGTAAG
AA242159 mouse CTTCCGCGCTTTTCATTACCGTACGCACCGGTCA-CGATCGGCATCGCGGAGGATCGGTCATGGGACTTTGCAAG
```

FIGURE 10

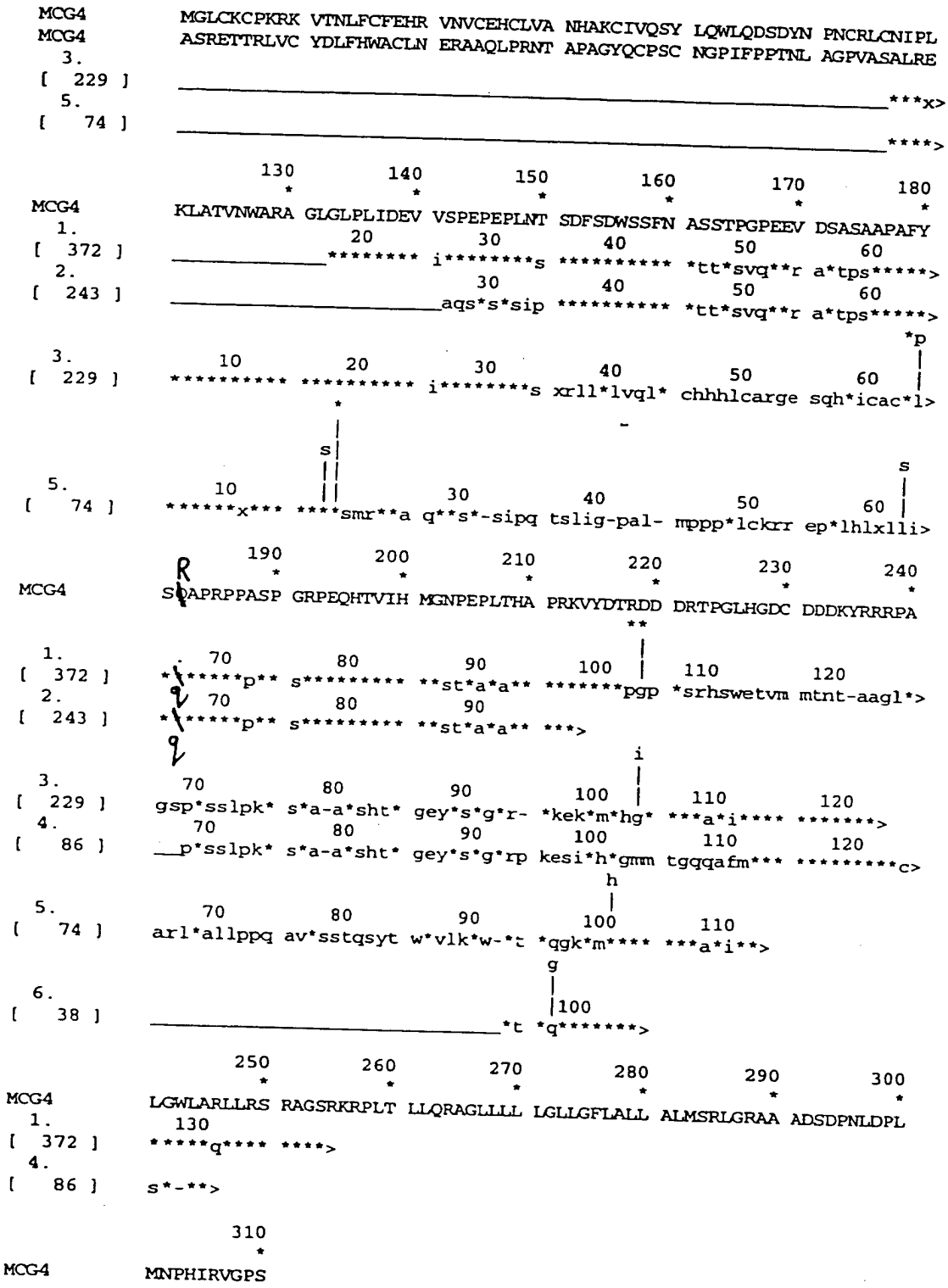

MCG4 MGLCKCPKRK VTNLFCFEHR VNVCEHCLVA NHAKCIVQSY LQWLQDSDYN PNCRLCNIPL 60
MCG4 ASRETTRLVC YDLFHWACLN ERAAQLPRNT APAGYQCPSC NGPIFPPTNL AGPVASALRE 120
3.
[229]
5.
[74]
MCG4 KLATVNWARA GLGLPLIDEV VSPEPEPLNT SDFSDWSSFN ASSTPGPEEV DSASAAPAFY
1.
[372]
2.
[243]
3.
[229]
5.
[74]
MCG4 SQAPRPPASP GRPEQHTVIH MGNPEPLTHA PRKVYDTRDD DRTPGLHGDC DDDKYRRRPA
1.
[372]
2.
[243]
3.
[229]
4.
[86]
5.
[74]
6.
[38]
MCG4 LGWLARLLRS RAGSRKRPLT LLQRAGLLLL LGLLGFLALL ALMSRLGRAA ADSDPNLDPL
1.
[372]
4.
[86]
MCG4 MNPHIRVGPS

Figure 10 (Continued)

Search Analysis for Sequence: MCG4

Search from 1 to 310

Date: September 22,1997

Matrix: pam250 matrix

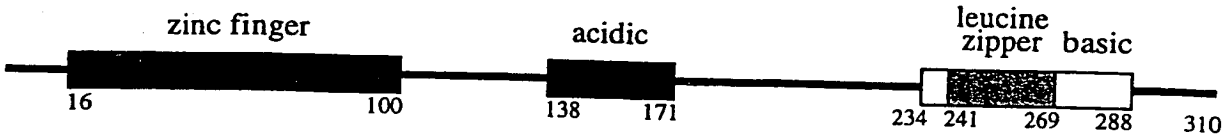
Score Region from 1 to 310

Maximum possible score: 1598

Aligned sequences:

1. = EST AA074703 phase 1 translation
2. = EST AA134788 phase 3 translation
3. = EST AA134788 phase 2 translation
4. = EST AA074703 phase 3 translation
5. = EST AA074703 phase 2 translation
6. = EST AA134788 phase 1 translation

FIGURE 11 Domains of MCG4



zinc finger consensus: $CX_2HX_4CX_2CX_4HX_2CX_{17}CX_2CX_{18}HX_2CX_{18}CX_2C$

acidic domain consensus: 9/34 negatively charged amino acids, 0/34 positively charged

basic domain consensus: 13/55 positively charged amino acids, 0/55 negatively charged

leucine zipper domain consensus: $LX_6LX_6RX_6LX_6L$

alternate "novel" leucine zipper-like motif where leucine would not be aligned along the one surface of an alpha helix domain: (aa 261) $LX_6LXLX_6LXLX_6L$ (aa 286)